Burst Image Super-Resolution with Base Frame Selection
– Supplementary Material –

Sanghyun Kim* Minjung Lee* Woohyeok Kim Deunsol Jung Jaesung Rim
Sunghyun Cho Minsu Cho

Pohang University of Science and Technology (POSTECH), South Korea

{sanghyun.kim, minjlee, woohyeok, deunsol.jung, jsrim123, s.cho, mscho}@postech.ac.kr

A. Appendix

In this supplementary material, we provide more details of synthetic-NEBI configurations in section A.1, and additional details of the post processing for real-NEBI in section A.2. In section A.3, we present more examples of syn-/real-NEBI dataset. In section A.4, we measure the test oracle of Frame Selection Network(FSN) to confirm how much performance improvement is possible. We also test our FSN in public benchmark in section A.5. In section A.6, we show additional qualitative results.

A.1. Detailed configurations of synthetic-NEBI

In this section, we explain the detailed configurations of generating the synthetic dataset, especially synthesizing motion blur and injecting noise. Following [1, 3], we convert images to RAW images through inverse gamma compression, inverse white balance, and inverse color conversion, and then synthesize motion blur and Poisson-Gaussian noise.

For the motion blur synthesis, we utilize gyro sensor data acquired from the Samsung Galaxy S22. Specifically, considering a burst of 14 images taken with different exposure times (varying from 0.01s to 0.14s, increasing at intervals of 0.01s), we randomly sample consecutive gyro sensor values matching each image's exposure. For example, we sample a sequence of 2 consecutive gyro measurements for exposure times such as 0.01s, 0.02s with 4 samples, ..., 0.14s with 28 samples. Subsequently, we interpolate these values eight-fold to create a smooth blur trajectory and then calculate homography. Finally, we obtain the blurred image by warping the ground truth sharp images corresponding to each burst using the previously derived homography and subsequently averaging them.

After synthesizing motion blur, we inject Poisson-Gaussian noise into the burst images. We model noise in an image as:

$$B_{noisy} = (B + N_{shot}) + N_{read}$$

where B is a blurred image from the previous step and B_{noisy} is a noise injected image. N_{shot} and N_{read} are shot noise and read noise, respectively. We use the Poisson noise and Gaussian noise model for the shot noise and read noise as follows:

$$(B + N_{shot}) \sim \mathcal{P}\left(\frac{B_{photon}}{\lambda_{shot}}\right)\lambda_{shot} \quad (1)$$

$$N_{read} \sim \mathcal{N}(0, \lambda_{read}) \quad (2)$$

where B_{photon} is the number of photons in the blurred image. λ_{shot} and λ_{read} are shot noise (*i.e.*, a mean of Poisson distribution $\mathcal{P}$) and read noise parameters (*i.e.*, a variance of Gaussian distribution $\mathcal{N}$), respectively. We calibrated shot and read noise parameters across a range of ISO settings, from ISO 100 to 12800.

To generate noises of arbitrary ISO values, we model the linear relationships of ISO values, shot noise, and read noise parameters, similar to [3]. We first randomly sample an ISO value of the shortest exposure frame (*i.e.*, 0.01 seconds in synthetic-NEBI) as:

$$\text{ISO}^{0.01} \sim \mathcal{U}(100, 12800) \quad (3)$$

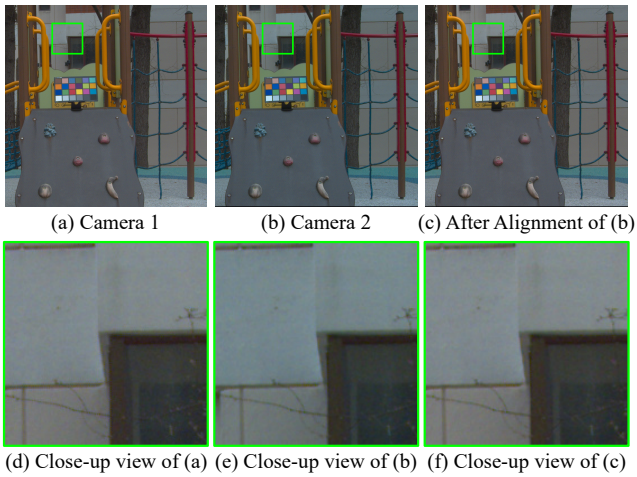
where $\mathcal{U}$ is a uniform distribution, and $\text{ISO}^{0.01}$ is an ISO value of the shortest exposure frame. According to the exposure ratio of the burst images, we compute the ISO values of other frames as follows:

$$\text{ISO}^{s} = \frac{0.01}{s} \cdot \text{ISO}^{0.01} \quad (4)$$

where $\forall s \in \{0.02, \ldots, 0.14\}$. ISO^{s} is an ISO value of the frame corresponding to the exposure s. Then, we use a linear model between the ISO values and the shot noise parameters as follows:

$$\hat{\lambda}^{s}_{shot} = 9.2857ee{-}07 \times \text{ISO}^{s} + 8.1006\text{e}{-}05 \quad (5)$$

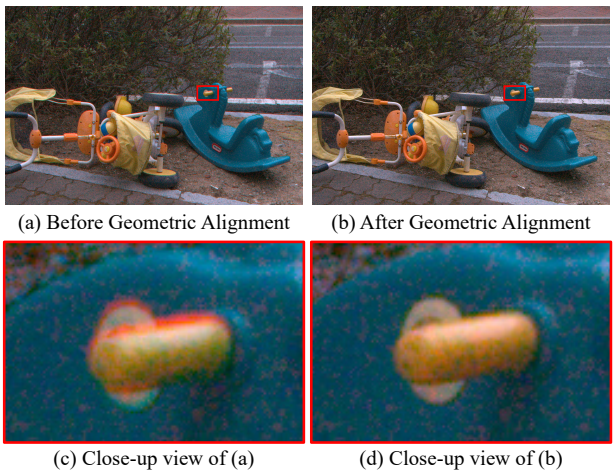
where $\hat{\lambda}^{s}_{shot}$ is a computed shot noise parameter of each frame in the synthesized burst images. The multiplier and



(a) Camera 1 (b) Camera 2 (c) After Alignment of (b)

(d) Close-up view of (a) (e) Close-up view of (b) (f) Close-up view of (c)

Figure 1. **Result of photometric alignment.** (a) and (b) are images captured with two camera modules. (c) shows the photometric alignment result of (b). As shown in (d) and (e), color differences exist before the alignment. (f) shows our photometric alignment significantly reduces the differences.



(a) Before Geometric Alignment (b) After Geometric Alignment

(c) Close-up view of (a) (d) Close-up view of (b)

Figure 2. **Result of geometric alignment.** (a) and (b) show stereo-anaglyph images, where the first burst frame and its corresponding ground-truth frame are visualized in red and cyan, respectively. (c) shows red and cyan lights are not aligned due to misalignment between the two frames. As shown in (d), red and cyan lights are better aligned after the geometric alignment.

intercept are estimated from the calibrated shot and read noise parameters using the linear regression.

We also estimated the linear relationship between the shot and read noise following [3]. We model read noise parameters as:

$$log(\hat{\lambda}^s_{read}) = 2.2282 \times \log(\hat{\lambda}^s_{shot}) + 0.45982 \qquad (6)$$

where $\hat{\lambda}^s_{read}$ is a read noise parameter of the frame in the burst images. Finally, for each set of burst images, we randomly sampled ISO values, shot, and read noise parameters and then synthesized noise on blurred burst images using Eq.(1) and Eq.(2).

With the proposed blur and noise synthesis, we synthesized non-uniformly exposed 14 burst frames, which have varying amounts of blur and noise. The first frame in the burst sequence has the shortest exposure, resulting in less blur but more noise. Conversely, the last frame exhibits more blur but is less noisy. The burst frames are 1-channel RAW images and the corresponding ground-truth images are 3-channel raw-RGB images without mosaicing applied. We utilize the sharp frames of the GoPro dataset [8] to synthesize our dataset. Finally, we synthesize a total of 2,750 burst sets, and each burst set consists of 14 burst frames and their corresponding ground-truth frames. The shape of each burst frame and ground-truth frame is $4 \times 80 \times 80$ and $3 \times 640 \times 640$, respectively.

A.2. Post-processing for real-NEBI

Despite the sophisticated hardware design of the dual-camera system, geometric misalignment may occur between the two camera modules. Due to the optical spectrum differences of the beam splitter and ND filter, photometric misalignment may also occur. So, we adopt post-processing to acquire more accurate ground truth images.

First, we demosaic 14 burst frames and ground-truth frames captured in the RAW format using a demosaicing algorithm of Malvar et al. [7]. Then, we apply photometric and geometric alignment to the ground truth images. For the photometric alignment, we capture a static scene with a color chart using two camera modules and obtain two reference images, as done in [9]. We estimate a 3×3 matrix that matches the values of the color chart captured with one of the camera modules to the other. The matrix is computed by solving a least-squares problem. Then, all ground-truth images are corrected using the estimated matrix. Figure 1(a)-(b) show noticeable color differences between images before photometric alignment. After the photometric alignment, these color differences are substantially reduced, as shown in Figure 1(c).

For the geometric alignment, we estimated the homography matrix from the first burst frame and the corresponding ground-truth frame. The two frames are simultaneously captured with the same exposure time and different gain values, so they have the same contents but different amounts of noise. We use the enhanced correlation coefficient method [5] for estimating the homography matrix, and we found the method performs well even with different amounts of noise. For each burst set, we estimated a homography matrix and applied the matrix to a ground truth image. Figure 2(d) shows a result of geometric alignment, where the red and cyan lights are better aligned after the geometric alignment.

After the alignment, the burst frames are down-sampled

Model	Frame selection	PSNR↑	SSIM↑	LPIPS↓
BIPNet [4]		33.774	0.920	0.108
	✓	**33.878**	**0.922**	**0.106**
	oracle	34.180	0.926	0.101
Deep-sr [1]		33.516	0.917	0.117
	✓	**33.872**	**0.921**	**0.107**
	oracle	34.140	0.924	0.099
Deep-rep [2]		34.059	0.925	0.103
	✓	**34.313**	**0.928**	**0.099**
	oracle	34.539	0.931	0.094

Table 1. **Oracle performance on synthetic-NEBI.** We train the models from scratch on the synthetic-NEBI. The term 'oracle' denotes the oracle performance, and '✓' denotes the results of the model with FSN. The best score is highlighted in bold. This result shows that FSN improves the enhancement quality.

Model	Frame selection	PSNR↑	SSIM↑	LPIPS↓
BIPNet [4]		33.957	0.904	0.143
	✓	**34.367**	**0.901**	**0.157**
	oracle	35.650	0.912	0.142
Deep-sr [1]		31.084	**0.908**	0.180
	✓	**31.149**	0.907	**0.180**
	oracle	33.095	0.926	0.163
Deep-rep [2]		31.089	**0.911**	**0.174**
	✓	**31.308**	0.904	0.194
	oracle	33.047	0.929	0.163

Table 2. **Oracle performance on real-NEBI.** We use the models pretrained on the synthetic-NEBI dataset. The term 'oracle' denotes the oracle performance, and '✓' denotes the results of the models with FSN. The best score is highlighted in bold. This result shows that FSN improves the enhancement quality.

(*i.e.*, $\times 1/4$) using linear interpolation for the input of the enhancement pipeline. The down-sampled burst frames are mosaicked and saved in the RGGB format. Finally, we collect a total of 96 burst sets, and each burst set consists of 14 burst frames and their corresponding 14 ground-truth frames. The shape of each burst frame and ground-truth frame is $4 \times 148 \times 238$ and $3 \times 1184 \times 1904$, respectively.

A.3. Additional synthetic-/real-NEBI visualization

Figures 3 and 4 present additional examples from the synthetic-NEBI and real-NEBI datasets, respectively. Both datasets consist of burst sequences with non-uniform exposure conditions. The frames with shorter exposure are likely to have more noise, while those with longer exposure are likely to exhibit more blur.

A.4. Oracle performance of Frame Selection Network

In Tables 1 and 2, we provide a comprehensive performance analysis, including the oracle performance, of the previous burst super-resolution networks, BIPNet [4], Deep-sr [1], and Deep-rep [2], on the synthetic-NEBI dataset and real-NEBI dataset. The oracle performance represents the average performance when only the optimal base frame is chosen. To obtain the oracle performance, we forward each of the 14 burst frames once as a base frame and then evaluate by comparison with its corresponding ground-truth. There is a substantial gap between the oracle performance and each baseline using individual existing models. This discrepancy underscores the pivotal role that the choice of the base frame plays in determining the quality of the final output. Our proposed approach, FSN, consistently surpasses the performance of each baseline.

A.5. Frame Selection Network on public benchmark

We also examine the impact of FSN on public benchmarks. Bhat *et al.* [1] provide a synthetic and a real burst dataset named SyntheticBurstSR and RealBurstSR, respectively. For BurstSR, we evaluate each baseline, the three existing burst super-resolution models [1, 2, 4] as the same as above, with and without our FSN. In this experiments, we set the number of CMA blocks, denoted as L, as 4. Table 3 demonstrates the impact of FSN on a public benchmark. FSN improves the performance compared to the baselines; however, the gain on the public benchmark is marginal than on the NEBI dataset. These differences may arise since the public benchmark captures scenes with uniform exposure time settings, whereas our NEBI dataset assumes dynamic exposure times. Figure 5 vividly illustrates the disparities between the public benchmark and NEBI dataset. In the synthetic-NEBI dataset, where we assume a non-uniform exposure setting, we utilize a video dataset [8] and synthesize gyro blur, making the motion more noticeable, while the SyntheticBurstSR [1] exhibits minimal motion. Additionally, the NEBI dataset features a wide range of noise levels, whereas SyntheticBurstSR maintains a narrow range of noise levels. For the real dataset, the real-NEBI dataset captures scenes with large motion and dynamic noise for simulating non-uniform exposures; RealBurstSR [1] exhibits small motion and similar noise levels between frames. Since the FSN is designed to consider motion information and the complementary characteristics of diverse exposure frames, the performance improvement may not be significant on the public benchmarks.

A.6. Additional Qualitative results

Figures 6 and 7 show qualitative results on the synthetic-NEBI and real-NEBI datasets, respectively. For example, in the second row of Figure 6, a man's eye appears closed and

Benchmark	Model	Frame selection	PSNR↑
SyntheticBurstSR [1]	BIPNet [4]		40.673
		✓	40.732
	Deep-sr [1]		38.330
		✓	38.341
	Deep-rep [2]		40.181
		✓	40.183
RealBurstSR [1]	BIPNet [4]		47.870
		✓	47.870
	Deep-sr [1]		47.699
		✓	47.737
	Deep-rep [2]		48.315
		✓	48.321

Table 3. **Results on public benchmark.** We train FSN on the provided pretrained model. '✓' denotes the results with the FSN and the proposed loss functions on top of the baseline model.

blurry in the original BIPNet result, while applying the FSN results in a clearer and more open eye. Similarly, in the third row, the woman on the right in the yellow box has some artifacts on her face in the Deep-sr results. However, when the FSN selects the base frame well, the artifacts disappear. In the last row, the woman in the yellow box has a blurry mouth in the results of Deep-rep. In contrast, with the addition of the FSN, the woman's mouth is clearer. In Figure 7, the FSN opts to choose a less noisy frame as the base. This strategic choice consistently yields clear output by integrating all frames well based on the chosen base frame.

A.7. Details about Auto-Exposure model

The common traditional approach for AE is to measure optimal exposure based on entropy. Following previous methods [10, 11], we select the base frame with the maximum entropy calculated based on the image's histogram. Specifically, the entropy e_i for i-th burst frame is calculated as follows:

$$n_b = \sum_{x \in H, y \in W} \delta(I_i(x, y) - b), \tag{7}$$

$$p(b) = \frac{n_b}{\sum_{j=0}^{B} n_j}, \tag{8}$$

$$e_i = -\sum_{j=0}^{B} p(j) \log p(j), \tag{9}$$

where δ is the discrete Dirac delta function, B is the number of bins of a histogram and $I(x, y)$ is the pixel intensity of the i-th image corresponding to the (x, y) coordinate. After calculating the entropy for all frames in burst shots, we select the frame with the highest entropy value as the base frame.

A.8. Implementation Details

Following the previous super-resolution methods [1, 2, 4] we use the number of the input burst frames as $N = 14$ and the scale factor $\times 4$. We train the super-resolution network using the first frame of the input burst frames as a baseframe. To acquire a ground-truth base frame index, we identified the frame in each burst that yields the highest performance on trained models. Specifically, for each burst, by designating each frame as the base frame, we generate predictions and compute the PSNR values against the corresponding ground-truth high-resolution image. The frame index associated with the highest PSNR value is determined as the ground-truth index. Our FSN uses $L = 2$, a batch size of 32, and the AdamW optimizer [6] for training.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Deep burst super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9209–9218, 2021. 1, 3, 4, 8, 9, 10

[2] Goutam Bhat, Martin Danelljan, Fisher Yu, Luc Van Gool, and Radu Timofte. Deep reparametrization of multi-frame super-resolution and denoising. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2460–2470, 2021. 3, 4, 9, 10

[3] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11036–11045, 2019. 1, 2

[4] Akshay Dudhane, Syed Waqas Zamir, Salman Khan, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Burst image restoration and enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5759–5768, 2022. 3, 4, 9, 10

[5] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE transactions on pattern analysis and machine intelligence*, 30(10):1858–1865, 2008. 2

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[7] Henrique S Malvar, Li-wei He, and Ross Cutler. High-quality linear interpolation for demosaicing of bayer-patterned color images. In *2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages iii–485. IEEE, 2004. 2

[8] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *CVPR*, 2017. 2, 3

[9] Jaesung Rim, Geonung Kim, Jungeon Kim, Junyong Lee, Seungyong Lee, and Sunghyun Cho. Realistic blur synthesis for learning image deblurring. *arXiv preprint arXiv:2202.08771*, 2022. 2

[10] SaiKiran Tedla, Beixuan Yang, and Michael S Brown. Examining autoexposure for challenging scenes. In *Proceedings*

of the IEEE/CVF International Conference on Computer Vision, pages 13076–13085, 2023. 4

[11] Chi Zhang, Zheng You, and Shijie Yu. An automatic exposure algorithm based on information entropy. In *Sixth International Symposium on Instrumentation and Control Technology: Signal Analysis, Measurement Theory, Photo-Electronic Technology, and Artificial Intelligence*, pages 152–156. SPIE, 2006. 4

Figure 3. **Visualization of the synthetic-NEBI dataset.** The left four columns display a subset of 14 input bursts, while the rightmost column presents the ground-truth image corresponding to the first input frame. From the left to the right in burst, the exposure time increases. For a better view, we visualize the images with the same size. Note that the burst frames simulate being shot with a gradually increasing exposure time.

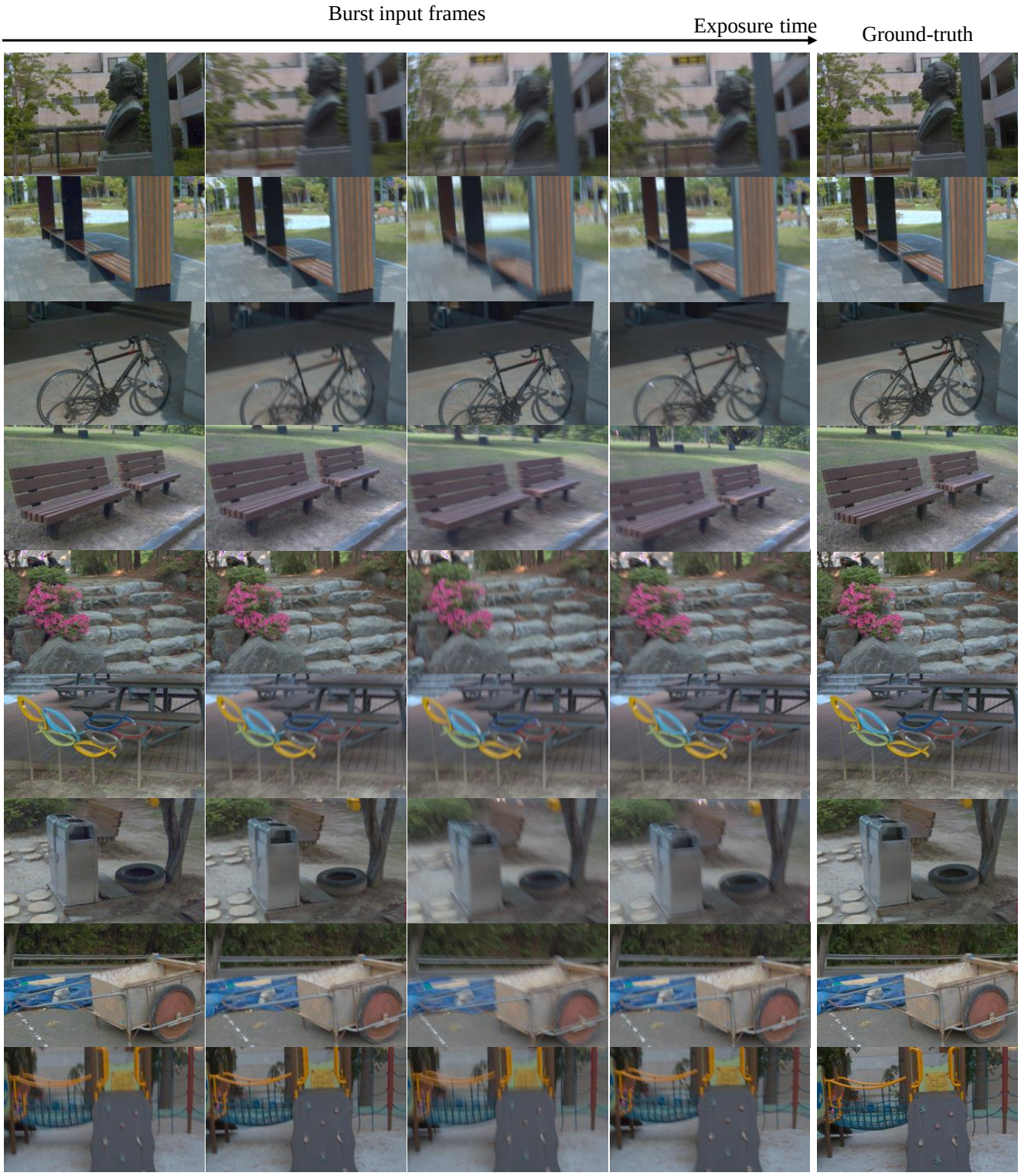


Figure 4. **Visualization of the real-NEBI dataset.** The left four columns display a subset of 14 input bursts, while the rightmost column presents the ground-truth image corresponding to the first input frame. From the left to the right in burst, the exposure time increases. For a better view, we visualize the images with the same size. Note that the burst frames simulate being shot with a gradually increasing exposure time.

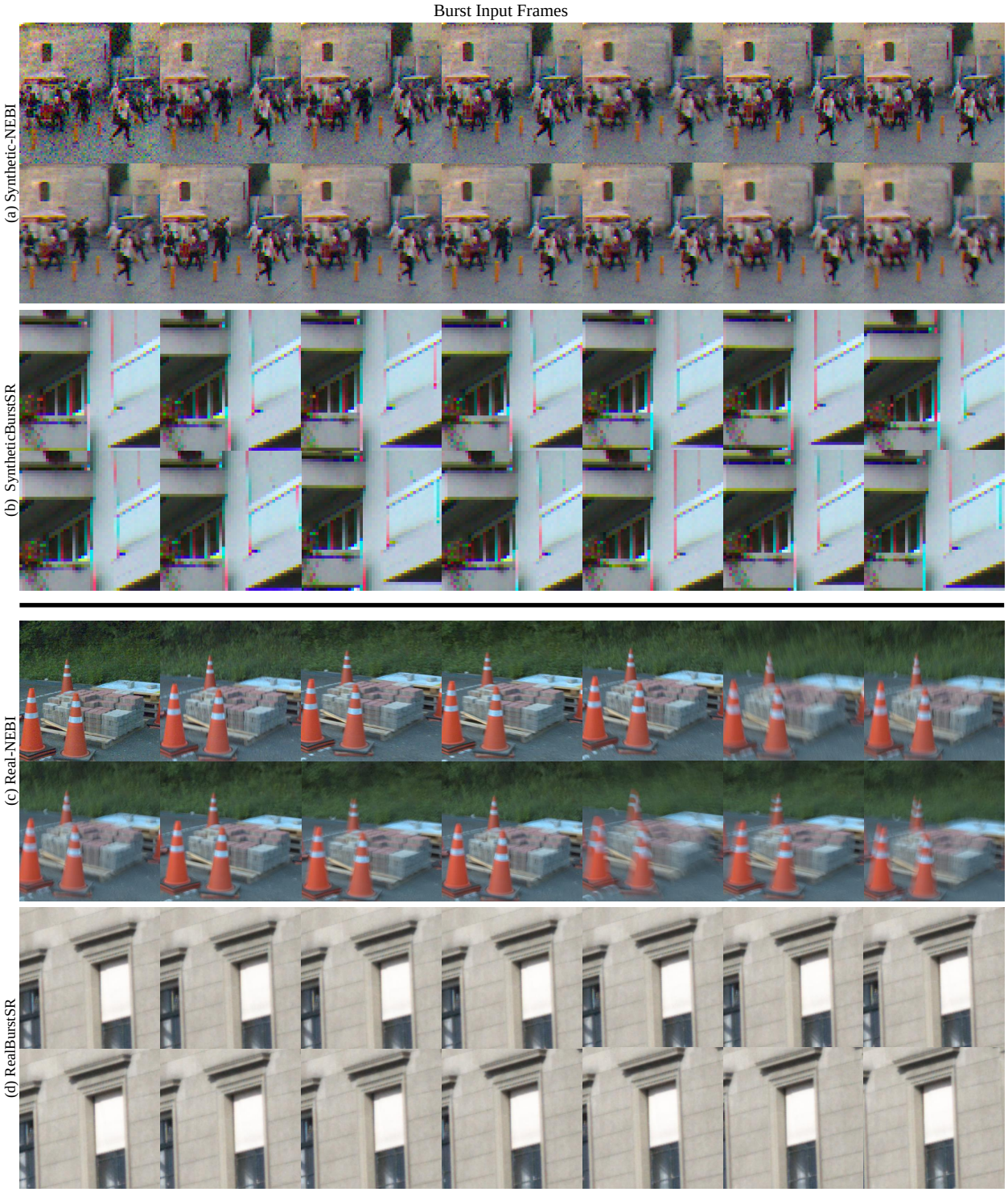


Figure 5. **Comparison of NEBI and public benchmark.** The two datasets (a) and (b) are synthetic datasets, and the two datasets (c) and (d) are real datasets. Comparing (a) and (b), synthetic-NEBI exhibits more dynamic motion, blur, and noise than SyntheticBurstSR [1]. Similarly, (c) and (d) highlight the differences between the two real datasets. The real-NEBI features significant motion and blur with dynamic noise, whereas RealBurstSR [1] has minimal motion and similar noise between frames.

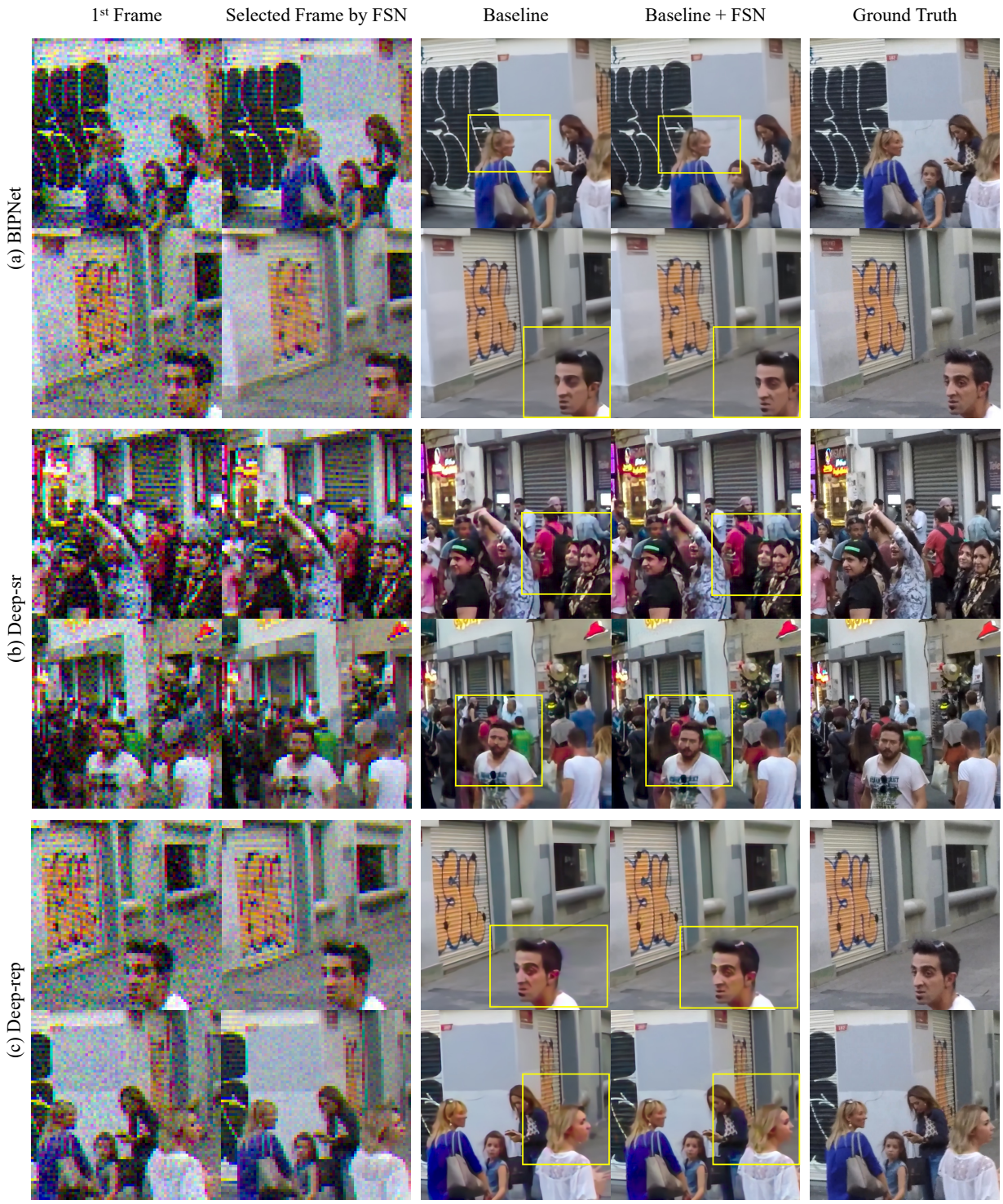


Figure 6. **Qualitative comparison on synthetic-NEBI.** We conduct a comparative analysis between the results obtained by applying three previous methods individually—(a) BIPNet [4], (b) Deep-sr [1], and (c) Deep-rep [2]—and the FSN. 'Baseline' denotes the results when the previous methods are applied individually, while 'Baseline + FSN' indicates the integration of the FSN on top of the baseline in a plug-and-play manner. The FSN demonstrates notable improvements by effectively merging complementary information from multiple frames, thereby enhancing high-frequency image details. In contrast, the existing burst models (a), (b), and (c) tend to predict images with artifacts or blur. Best viewed when zoomed in.

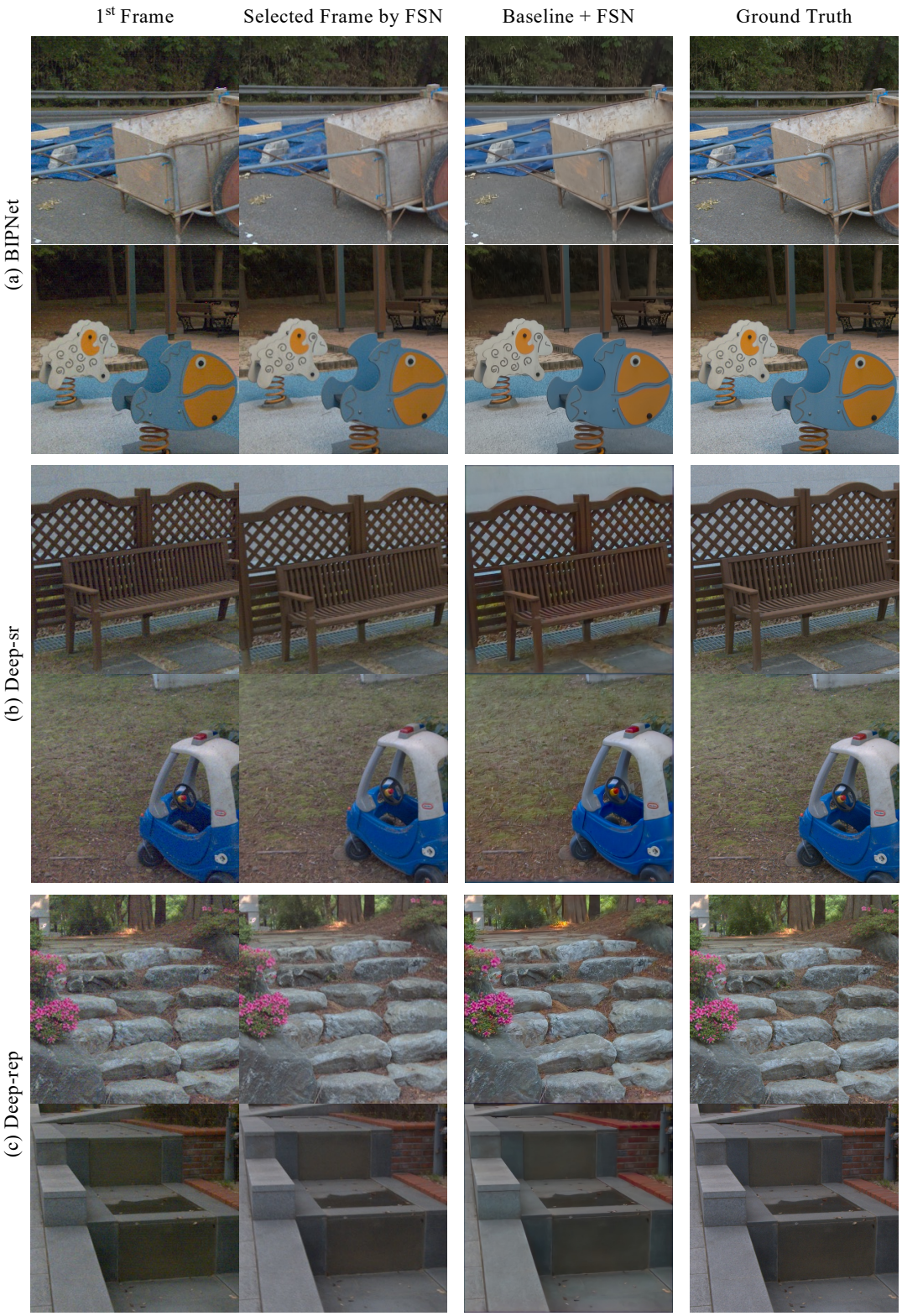


Figure 7. **Qualitative comparison on synthetic-NEBI.** We present the results of three previous methods—(a) BIPNet [4], (b) Deep-sr [1], and (c) Deep-rep [2]—with the FSN. 'Baseline + FSN' indicates the integration of the FSN on top of the baseline in a plug-and-play manner. The FSN selects a frame with less noise than the 1st frame well, resulting in improved visual quality. Best viewed when zoomed in.